DOCUMENT RESUME

ED 163 069 TM 008 151

AUTHOR Smith, Laura Spooner
TITLE Decisions and Dilemmas in Constructing Criterion-Referenced Tests: Some Questions and Issues. CSE Report No. 110.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Mar 78
NOTE 29p.

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Behavioral Objectives; *Criterion Referenced Tests; Educational Objectives; State of the Art Reviews; *Test Construction; *Testing Problems; Test Interpretation; Test Items

ABSTRACT

Because criterion referenced measurement is a relatively new approach to testing, systematic guidelines for constructing criterion referenced tests (CRTs) have not yet been developed. A number of questions and issues face the test developer who wishes to provide a test that will be helpful in a particular decision-making situation. There are three basic steps in the generation of domain specifications guiding CRT construction. First, decide what the test should measure. The instructional independence of the test items refers to the items' dependence on or relationship to the instructional curriculum content. The amount of content measured by a domain may be conceptualized as formative or summative evaluation. Secondly, determine the dimensions of the class of behaviors (or performance) to be tested. Tasks and responses must be selected which indicate that the learner has demonstrated achievement or competence; this may be the most difficult aspect of test construction. The role of items which measure transfer of training must also be determined. Finally, develop a set of item generation guidelines. Items should be suitable for inclusion in an item bank. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Center for the Study of Evaluation

UCLA Graduate School of Education
Los Angeles, California 90024

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

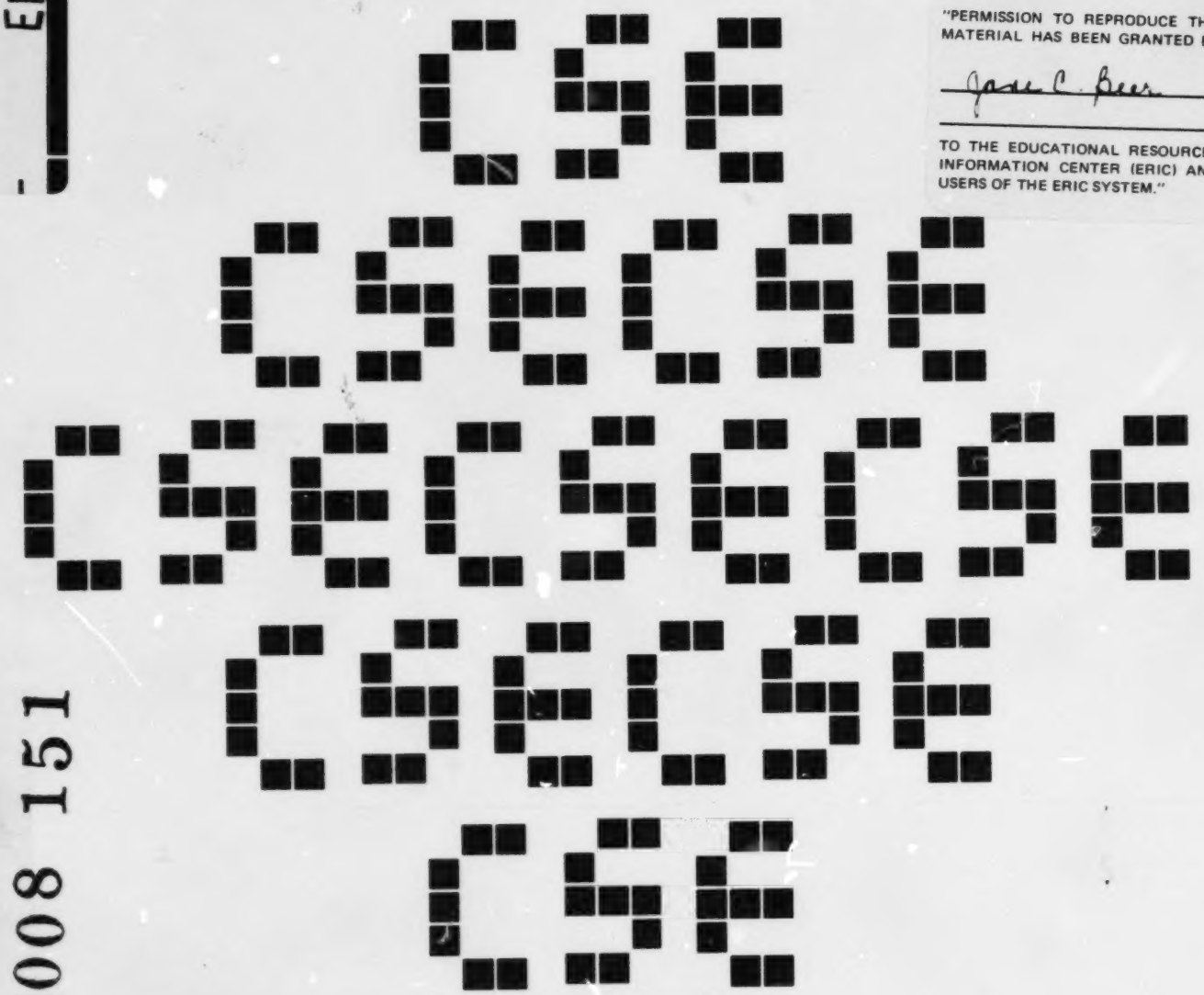
"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Jane C. Beer

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

ED163069

TM008 151



Center for the Study of Evaluation

The Center for the Study of Evaluation (CSE) is a university-based research and development center established in 1966 by the Department of Health, Education, and Welfare. The mission of the Center has been to conduct research, development, training, and dissemination activities for the purpose of improving the planning and evaluation of education and other social action programs. More specifically, CSE activities include: guidance for educators and researchers in the areas of tests and measures and in educational decision making; development of easy to use evaluation procedures and materials in the form of manuals, workshops and kits; and evaluation of projects for federal, state, local, and university agencies requesting technical assistance in evaluation. The work of CSE during its first decade has clarified the ways in which systematic evaluation can be helpful in the improvement of many kinds of programs.

Information about CSE and its publications may be obtained by writing to:

Director, Public Information
Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California 90024

DECISIONS AND DILEMMAS IN CONSTRUCTING CRITERION-REFERENCED TESTS: SOME QUESTIONS AND ISSUES

Laura Spooner Smith

CSE Report No. 110

March 1978

Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California 90024

The research reported herein was supported in whole or in part by the National Institute of Education, Department of Health, Education, and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

DECISIONS AND DILEMMAS IN CONSTRUCTING CRITERION-REFERENCED TESTS: SOME QUESTIONS AND ISSUES

Laura Spooner Smith

FOREWORD

In the field of criterion-referenced measurement, it may be both the best of times and the worst of times. With energy and enthusiasm, educators, both practioners and researchers, have been addressing difficult topics within this relatively young area of measurement. Indeed, Klein and Kosecoff (1973) noted that we are likely to "witness during the next few years a number of major contributions to testing theory and methodology arising from the use of criterion-referenced testing." Some have high expectations that the efforts of those concerned with criterion-referenced testing will lead to advances in instructional technology.

For those of us who are looking for reasonable and generalizable principles to guide the construction and use of criterion-referenced tests (CRT's) it may, however, be the worst of times. Why are there so few applied guidelines in the area of CR measurement? One explanation lies with the relative "youth" of the technology; as in most young fields, research efforts are somewhat fragmented and uneven. While some researchers have conducted empirical studies on different facets of CRT construction (e.g., E. L. Baker, 1971; Dahl, 1971), some have investigated statistical properties of the tests (e.g., Harris, 1975). Still others have proposed analytic schemes for generating good CRT's (e.g., Hively, 1973; Bormuth, 1970). Millman (1974) has deftly

[1]This paper was written in Winter 1976 as part of graduate study in the UCLA Graduate School of Education.

summarized the state of the art by noting that "...the concept [of CR measurement] has gained greater acceptance as a type of general interpretation... than it has as a testing procedure."

A slightly more subtle, but equally exasperating, quality of CR measurement in the 1970's is the diversity of purposes and settings to which the technology has been applied. CRT's have been used, for example, for research in curriculum development (e.g., Hively, 1973); large-scale assessment (e.g., Wilson, 1974); computer assisted evaluation systems (such as the Comprehensive Achievement Monitoring system used by the New York State Department of Education); and classroom management (e.g., Skager, 1974). The benefit of this diversity has been the opportunity to observe, document, and consider the relative merits of different CRT efforts; the bane is the difficulty of generalizing across instances of use in order to isolate the critical attributes of "good" CRT construction procedures.

Popham (1975) called for a "programmatic effort to sharpen our criterion-referenced methodology." Until an adequate methodological base for CR measurement is established, Popham further noted, developers of CRT's will continue to face "tensions and terrors" as they respond to the rising call for suitable criterion-referenced measures.

The experience of writing this paper suggests that those who venture into the relatively uncharted territory of CR measurement are likely to find more questions than answers. Consequently, the intent of this document is neither to advocate what <u>should be</u> in the field of criterion-referenced testing, nor to describe precisely what <u>is</u>. Rather, what follows represents an effort to pinpoint some practical problems in CRT development by identifying plausible--not necessarily preferable--alternative procedures that we are apt to encounter along the development path.

The term "domain specification" is commonly used to denote the set of specifications that guides the construction of a criterion-referenced measure. While at present there is no universally accepted method for developing domain specifications, many leaders in the field of CR measurement (Popham, 1975; E.L. Baker, 1974; Hively, 1973; Millman, 1974) agree that a domain specification ought to include at least the following elements: 1) a description of the objectives, "chunks" of content, or class of behaviors that the test is intended to measure; 2) specification of the dimensions of performance, e.g., the stimulus and response properties of test items or tasks; and 3) a set of rules or directions to guide the construction of items. Test items, it should be noted, need not be restricted simply to pencil and paper measures of achievement.

In addition, there appears to be a logical sequence of steps to follow when generating a domain specification. Nitko (1974), however, noted that in practice the sequence is rarely followed and quite often the developmental procedure is both iterative and intuitive. Nonetheless, procedures that a CRT developer follows can be roughly categorized into the following three "steps:"

1. Deciding what the test should measure.
2. Determining the dimensions of the class of behaviors (or performance) to be tested.
3. Developing a set of item generation guidelines.

Potential developers of domain specifications may be encouraged by the apparent simplicity of these three steps. All they need, it would seem, are a few more guidelines--perhaps an exemplary domain specification or two--and presto: they are ready to "spin out a semantic recipe" (Popham, 1975) for constructing a good criterion-referenced measure.

Eager CRT developers will, however, find their efforts immediately checked. Not only do we lack a set of replicable rules for generating domain specifications, but there also is considerable variation among test development specialists concerning specific procedures to follow within each developmental step. The problem is further confounded by the fact that "...the method chosen to resolve a particular issue at one stage in the development of a CRT is likely to have ramifications for other stages in the developmental process as well as in the interpretation of the scores obtained" (Klein & Kosecoff, 1973).

Faced with the charge of constructing a criterion-referenced test what, then, are educators to do? What approaches have already been tried and what assumptions underlie their use? A conscientious attempt to answer these questions, giving adequate attention to the many views and procedures advocated in CR measurement today, would be worthy of a volume (or two) of writing, much of which would be highly speculative. Even if such an ambitious manuscript were available, the magnitude of issues and options would likely overwhelm an educators who must begin to construct a CRT tomorrow. It is thus to practical decisions and dilemmas confronting CRT developers that this paper is directed.

Before turning to more substantive issues, a word about the assumptions and organization of this paper is in order. First, it is assumed that a major reason for applying CRT techniques within an educational setting is to make instructional decisions (Glaser, 1963). Moreover, the nature of the decision will vary along such dimensions as the characteristics of the instructional system (e.g., classroom, statewide program); the persons who will make the decision (e.g, teacher, principal, legislator); and the consequences of the decision (e.g., monitoring individual student mastery, modification of instructional techniques, expanding or dropping a program). It is hard to imagine a single, invariable set of domain specification guidelines that will result in CRT's to suit all of the decisions all of the time.

Next, it is assumed that an important responsibility of CRT developers is to become familiar with the decision context within which CRT's will be used. This is particularly critical when the developers are not part of the unit, such as a school district or state education agency, which commissioned them to develop the criterion-referenced tests. For example, a CRT assessing only desired end-of-program outcomes would not be especially appropriate for quick turnaround formative evaluation purposes.

From these two assumptions springs the organizing rubric of this paper. For each of the developmental steps listed earlier, one or two questions--which CRT developers could pose to decision makers or to themselves--are presented. The questions represent my efforts to identify strategic points during the developmental process where fairly well recognized views in CR technology are apt to diverge. In addition, I have attempted to formulate

a limited number of questions that have direct implications for procedural, rather than theoretical, aspects of domain specification. A procedure could, of course, be considered the operational child of theories or assumptions. But, imagining myself in the position of an educator called upon to create a CRT for a very "real" school or district, I have shied away from substantive treatment of such tantalizing and critical topics as theories of performance, content analysis, and so on.

Finally, the reader will find, following each posed question, a brief synopsis of "options" open to CRT constructors. Here, again, the attempt has been made to isolate one or two recognized procedures, preferably those that have been applied or at least reasonably well explicated in the literature on criterion-referenced measurement. Whenever possible, studies or examples illustrative of a particular technique or approach are referenced.

STEP ONE: DECIDING WHAT THE TEST SHOULD MEASURE

In developing any type of achievement measure, an initial consideration is to determine what the test is to assess. For CRT developers, too, a preliminary decision is to select from an area (e.g., an instructional sequence), the content, behaviors, or objectives that are to serve as domains which test items or tasks will be designed to measure. Developers of CRT's may find the universe of possible "target" objectives considerably narrowed by attending to the following issues...

Question: How instructionally dependent should the CRT be? Instructional dependence refers to the extent to which a test item (or domain) is dependent upon or related to a particular curriculum, set of instructional materials,

or teaching technique. Surprisingly, there is no general agreement regarding the degree to which a CR measure should be directly referenced to a defined instructional sequence, or even whether performance dimensions should have a fixed relationship to an instructional program at all (i.e., educators who "yearn for ultimate criteria;" see Popham, 1975, pg. 132).

An advantage of instructionally dependent domains is that they are more likely to be "program fair" than are those derived from abstract analyses of content or behavior, or from those exclusively representing judgments regarding the "worth" of different objectives. Robert Baker (1974) contended that tests which are coupled to texts and instruction meet all of the requirements of the psychometric concept of content validity, while also contributing to the "definition of the behavioral structure of the subject matter domain treated." On the other hand, Skager (1975) noted that tests rooted in actual instructional content may fail to meet the validity criterion of "educational importance" as described by Cronbach (1969). While tests whose content directly reflect instructional events might be especially useful for monitoring the effectiveness of an instructional sequence during its developmental stages, i.e., during formative evaluation, the possible idiosyncratic nature of the domains may make evaluative comparisons between programs quite difficult.

Two contrasting views, both from the "instructional dependence school," are briefly discussed below. Following the discussion is an overview of alternate modes for determining test content, which are not necessarily sensitive to instructional intentions or transactions.

Instructionally dependent modes. An example of one extreme of instructional dependence are CR measures developed at the Southwest Regional Laboratory (SWRL), where "...all assessment procedures are systematically referenced to the particular objectives of the [instructional] program" (R. Baker, 1974). Not only are SWRL tests based on prespecified instructional outcomes, they also limit the population of learner responses to those called for during an instructional unit.

It should be noted that SWRL's approach to selecting domains constitutes an important link within an applied instructional management system which permits the classroom teacher to determine the extent to which specific outcomes are attained by individual students after instruction. While a critical precondition for constructing SWRL's tests are sets of intended instructional outcome statements,it also is possible to systematically infer the "architecture" of extant materials (see Smith, 1972).

Like Baker, Bormuth (1970) proposed that the content of achievement tests ought to be derived directly from instruction. Unlike the SWRL approach, however, Bormuth contended that it is possible to tie a test item to instruction without the intermediary of prespecified instructional objectives. Supporting Bormuth's approach, which utilizes linguistic principles to derive various item transformations from instructional content, is the notion that "...evaluation [of instruction] is not merely to find out something as narrowly limited as how well the instruction accomplishes the program designer's objectives." He argued that tests stemming from statements of intended outcomes may ignore possible side effects of

instruction, either positive or negative, and that, in fact, instructional programs may contain large amounts of instruction that are not specified in the objectives.

Analytic schemes. This approach to domain selection involves analysis of a subject matter area with the aim of identifying meaningful units of content, plus "performance" behaviors within an identified unit (Glaser & Nitko, 1971). Application of an analytic scheme may be important in cases where the curriculum itself is based upon an hypothesized taxonomy, hierarchy, or other theory. Hively's MINNEMAST Project (1973) serves as an instance in which CR measures were employed not only to assess learner progress, but also to validate certain hypotheses from which elements of the curriculum sprung. Hively described strategies for carrying out "analytic dialogues" between test writers and curriculum designers in order to arrive at a mutually agreed upon set of domains. Similar to the SWRL approach to domain selection discussed earlier, the MINNEMAST CRT's reflect the intentions of curriculum designers and, in this sense, could be viewed as "instructionally dependent."

Judgmental approaches. As implied by the name, judgmental approaches involve the judgments of groups or individuals in assessing the value or worth of classes of content or behaviors within an area of learning. The three examples which follow illustrate how judgmental approaches can be used to identify important domains for CR measures.

Popham (1974) reported on methods used at the Instructional Objectives Exchange (IOX), where test developers begin by mapping out "eligible contenders" for categories within a subject matter field, a procedure which

is accompanied by formal and informal advice from subject matter specialists. The IOX system constitutes an "item bank" from which educators can order domains and sets of related items. As such, the assumption apparently operating is that teachers, or other instructional decision makers, will request domain "sets" that are instructionally relevant to their own endeavors.

CR developers for the National Assessment of Educational Progress (NAEP) follow a procedure that entails reviewing recent literature to identify trends in a subject area. Existing sets of written objectives are then examined, followed by reviews by different panels of judges (Wilson, 1974). It is important to note that the NAEP effort is not targeted for a specific instructional system; rather, it is intended for use under a variety of instructional conditions. The loss of "instructional dependence" may, in fact, be compensated by the broad base coverage of content areas that result from the use of judgmental approaches.

Worth mentioning also are test content areas determined through "consensus" judgments. This technique typically involves surveying the opinions of various groups such as teachers, parents, curriculum experts, community members, in order to identify high priority items of curriculum. An applied example of such procedures is found in the CSE Needs Assessment Kit (Hoepfner, 1972), which provides directions and materials for systematically determining educational needs. Such procedures have clear limitations for developing measures sensitive to on-going instruction and may be most useful in "needs assessment" efforts for which, of course, criterion measures could be employed to determine learners' current status.

Question: Should a domain represent "terminal" or "enroute" behaviors? "Terminality," "domain size," "degree of generality," "levels of achievement," and other such terms refer to the same problem; how much instructional or subject matter territory should a domain encompass? A domain definition can range in "size" from, as Nitko (1974) suggested, the "...desired outcomes of an entire educational enterprise...to the specification of outcomes at the termination of a particular course." For practical-minded CRT developers, one helpful rule of thumb is to invoke Scriven's (1967) distinction between "formative" and "summative" type evaluations.

CRT's for formative and summative evaluations. The specific aspects of a formative type evaluation will vary according to size, length, and other characteristics of a program, but generally formative evaluation procedures are implemented in order to improve or revise an instructional sequence during its developmental stages. To be of practical value for instructional improvement, then, domains ought to be referenced to a meaningful unit of instruction that realistically can be modified.

Keesling (1975) asserted that formative evaluation of pupil progress aims to locate a student along a hierarchy of achievement. The intent, according to Keesling, is to assess the effectiveness of the curriculum in promoting mastery, while also evaluating the validity of an operationalized "taxonomic hierarchy." To this end, we might use Gagné's work on learning hierarchies (Gagné & Paradise, 1961), in which requisite or component, i.e., "en route," behaviors for a desired instructional objective are identified.

In contrast, Popham (1972) suggested: "If there is a degree of possible hierarchy present in the contending types of learner behavior... the chosen objective should represent the most terminal learner behavior." Popham's assertion makes clear sense for defining the domain "size" for a CRT used for end-of-program, or summative, evaluation efforts. The intent is definitely not to validate an hypothesized hierarchy of learning.

Significance of the behavior. Eva Baker (1974) offered a convincing argument against the use of domains with a magnitude of simple item equivalence. She warned that item equivalent tests, i.e., those based on objectives typical of the sixties behavioral objective movement, are likely to result in over-testing on relatively trivial aspects of instruction. Baker's precept, along with Popham's (1975) exhortation to identify behaviors that require at least a week to promote, may help CR test developers decide how large a "chunk" a domain ought to be.

STEP TWO: DETERMINING THE DIMENSIONS OF THE CLASS OF BEHAVIORS (OR PERFORMANCE) TO BE TESTED.

Probably the single most important attribute of a criterion-referenced measure is its power to yield unambiguous information about a learner's performance within a clearly defined class of behaviors. To achieve this type of clarity, domain specifications must precisely circumscribe the stimulus presented to learners as well as the form which their response will take (Cronbach, 1971). Before the stimulus and response properties of a domain can be stipulated, however, CR test developers must first select "tasks" that learners must perform in order to demonstrate a level of achievement.

Determining meaningful test tasks (or items) to measure a domain is perhaps the most difficult aspect of CR test construction. As Glaser and Resnick (1972) have pointed out, determining critical attributes of what is learned is not a well developed science, especially for complex behaviors. Indeed, prominent researchers in psychology and education have argued that failure to adequately perform a particular task does not necessarily signal lack of competence. Cole and Brunner (1971) have persuasively discussed this topic in relationship to linguistic abilities, with particular emphasis on groups that differ from the dominant middle class population. Bortner and Birch (1970) also have described how "underlying" competencies can be called forth by increasing or reducing stimulus competition within a task, noting that expressions of competency can vary along such dimensions as age, state of motivation, previous experience. Moreover, within a single well-defined performance domain, such as a precisely defined objective in mathematics, there exists a potential item pool of well over several thousand items (Hively, 1970).

The preceding observations have been included as testimony to the exasperatingly difficult decision CRT developers face when selecting specific performance tasks to measure achievement. Nonetheless, we might begin the selection process by considering the following question:

Question: To what extent should test-events (items or tasks) generalize within and beyond specified instruction? Answers to this question can range along a continuum from total reliance on "transfer" type tasks to those which are completely embedded in the actual stimulus and response properties of instruction.

Tasks derived from instruction. As an example of this approach we turn again to CRT's developed at SWRL, as described by R. Baker (1974). According to Baker, SWRL tests are rooted in a set of "instructional specifications" (IS) which map out both the instruction and its assessment. Mastery items are derived from the IS and serve the twin functions of practice (and hence could be considered part of instruction) and assessment. The job of the domain specifier is, then, to analyze the properties of the instruction in order to arrive at test tasks consistent with the stimuli presented and the responses practiced.

"Transfer" tasks. In marked contrast to instructionally embedded tasks is the position of E. L. Baker (1974) who urged developers of CRT's to identify transferable testing tasks. The issue at stake, according to Baker, is to provide domain specifications that maximize the range of instances and applications of a discrete learning objective. Baker sees domain specifications as an important tool for instruction, and they should be tools which promote the transfer of training. To illustrate her point, Baker noted: "The ability to list three causes for the depression could only be a suitable objective if, in domain context, it were modified to concern generalizable causes of economic decline, of which the 1929 depression was only one example."

Popham, who also advocates the selection of test tasks with maximum generalizability within and without a domain, offered an empirical method for validating the generalizability of test tasks. These procedures, which are reported in his book Educational Evaluation (1975), are intended to provided CR test developers with a sound rationale for selecting certain test tasks from among competing tasks.

A final note on "transfer" tasks is in order. Educators typically assume that the knowledge and abilities which a learner acquires in school have a strong relationship to "out-of-classroom" behaviors. However, the relationship between what learners are taught in school and how they actually perform in society is not clear. For this reason, researchers such as Popham (1975) and Nitko (1974) have urged CRT developers to identify tasks representing "proximate goals." Proximate goals, which define behaviors that a learner is expected to display at the end of a particular instructional sequence, need not directly reflect the materials and content of instruction, or even instances of practice. The aim is, rather, to identify significant behaviors that learners are expected to acquire as a result of instruction, but which were not necessarily rehearsed during instruction.

Question: What should constitute the stimulus and response attributes of a test task? Once a testing task has been selected, it is imperative that the CRT developer specify the stimulus properties of the task, as well as the response options available to the learner. It is from specifications of the stimulus and response characteristics of a domain that specific items will be developed.

Millman's (1974) assertion that a performance domain should be defined "...by those facets and elements that make a difference in how the learner responds" leads naturally to the truly baffling question, "What facets make a difference?" Millman tells us that our "intuition will help" --an answer which, unfortunately, captures the state of the art.

Some attempts have been made to identify "item format" characteristics that may affect performance. Hively (1973) has presented a "rudimentary

list" of such characteristics, including considerations such as the forms of instructions and responses, syntax, vocabulary, physical qualities of the materials. What we still do not know, however, is how such attributes are likely to influence a learner's response.

Though definitive answers to the question of critical "facets" are not even on the horizon, the problem cannot be overlooked by CRT constructors. What follows, then, is a brief outline of considerations to which CRT developers might wish to attend.

Instructional history. Harris (1974) suggested that tests, for the most part, are "designed to detect the experimental history of the student, and in that sense every test has an 'instructional bias'." The extent to which a CRT developer delimits stimulus and response properties of a behavior domain on the basis of instruction is, naturally, a function of the test's purpose. The case for attending to the instructional history of intended respondents was well made by Bormuth (1970):

> For one person, the item 9 X 12 = ___?___ may require only a rote memory process, for another a computational process, and for still another a count process, depending upon their respective instruction in mathematics. Explicitly, one of the several variables determining what process a given item tests is the relationship between the item and the instruction given the persons tested with the item.

The implication is that test items introducing content or behaviors not directly treated in instruction will detract from the purpose of administering the measure--[illegible] the extent to which specific properties of instruction [illegible]

Test bias. "Test bias" has been defined as a "group by test interaction" in which a group does not have the same shaped profile of scores across various tests being considered (Cleary, 1966). The notion of test bias has received increasing attention in recent years, particularly in the context of intelligence testing and with respect to cultural and linguistic biases. Though it is beyond the scope of this paper to examine sources of test bias, suffice it to say that CRT developers ought to attend to characteristics of target respondents, particularly when respondent groups are apt to be varied in terms of age, culture, or language experience.

Theoretical and empirical studies in content fields and related behaviors. Developers of CRT's would greatly benefit from results derived from empirical research on the content and process variables that affect performance. For example, a typical problem in specifying a domain is to delimit the classes of behavior (or content) from which distractors for selected response items are drawn. Davis and Diamond (1974) asserted that "effective distractors...should include natural misconceptions." To illustrate this point, they suggested that an item for college students to test knowledge of the word "pedantic" should include a distractor like "having feet" or "footed," because examinees might rely on irrelevant information (i.e., "pedes" is Latin for feet) to answer the item.

Davis and Diamond may have offered sound advise. But until solid research findings tell us more about processes underlying performance, the identification of "natural misconceptions" is entirely left to the domain

specifier's "best guess." "Best guesses" are not only likely to vary from individual to individual, but they are also difficult to explicate in clear, unambiguous terms. Simply directing an item writer to produce distractors that represent "reasonable incorrect answers" is not precise enough. Until research findings are available, CRT developer's are left to rely on their own perspicacity, the judgments of instructional and curriculum experts, plus empirical item try-outs.

STEP THREE: DEVELOPING A SET OF ITEM GENERATION GUIDELINES.

Educators administer criterion-referenced tests in order to have information about learners' status within a well-defined domain of behavior. If the domain has been clearly explicated, then we know what the test designers intended to measure. To have confidence in test results, though, we must also be confident that the test items are, in fact, consistent with the domain description. The job of the item writers is, then, to generate items that adhere to the specifications imposed by the domain definition; the job of the domain definer is to provide guidelines to insure such adherence.

One generally agreed upon characteristic of CRT items should be mentioned at this point. This is the notion that the sample of items included on the test are representative of a larger "pool" of items that could potentially be used to determine achievement of a domain. Given that the "pool" of acceptable items has been carefully delineated, then it follows that those items which are actually generated should realize a high degree of "definitional homogeneity" (Harris, 1974). Whether response patterns to such items should, in fact, display homogeneity is a matter open to debate (e.g., Millman, 1974; Bormuth, 1970).

To developers of CRT's one issue in item generation can be summed up in the following query:

Question: What method of test specification is likely to promote definitional homogeneity among items? Item generation schemes involve some logical, systematic, and replicable means for constructing items representative of the defined domain. Typical practice is to specify the task, with content replacements being variable, rather than identifying content and varying tasks (Harris, 1974). With the exception of computer item generation programs, such as that used by Millman at Cornell, item generation guidelines are prepared in some written form. Two examples of "formal item generation" approaches are briefly described below.

Item forms. Many CRT item construction approaches spring from the "item form" procedure developed by Hively and his associates for the MINNEMAST Project. The methods (and assumptions) involved in developing item forms have been thoroughly treated in CR literature, with Hively's own explanation perhaps the most widely cited (Hively, 1973). Statements about the objectives of the MINNEMAST curriculum, elicited from the curriculum designers, form the springboard for an item form. A completed item form would include the following elements: 1) general description of the task; 2) statements about the characteristics of the stimulus and response; 3) one or more "item form cells" specifying each class of items in the domain; and 4) the "item form shell" which provides rules for developing item variations from the one or more "replacement sets" of stimulus elements. The final result is a rather detailed scheme which is capable of objectifying item generation to the point where item writers, working independently, ought to be able to produce parallel tests.

Hively's approach, though, does have limitations. The first, and perhaps most obvious, is the amount of time and expertise required to create a significant number of item forms. Next, the sheer number of details included in a typical item form has the potential of rendering item writing a tedious and time consuming process. Patience would be an especially important virtue for item writers working with item forms. Finally, the technical complexity of an item form is likely to reduce its semantic utility for actually describing what the CRT purports to measure. Recall that one of the distinguishing features of a good CRT is its "descriptive power," that is, its ability to convey clear information about the status of a learner within a precisely delimited area of learning. If persons who are to make instructional decisions on the basis of CRT scores are themselves uncertain about what the test measures, then a major purpose for using criterion-referenced techniques has been defeated.

Amplified objectives. Popham (1975) directly addressed this problem in item generation schemes when he spoke of the trade-off between brevity (and interpretability) and precision. As a tentative and workable solution, he has described the characteristics of "amplified objectives," in use at the Instructional Objectives Exchange. Components of an "amplified objective" include, quite basically: 1) stimulus elements; 2) learner response options, and 3) a response format.

It should be clear from earlier discussions that whatever form item generation guidelines take, the final result reflects decisions made as the domain itself was conceptualized and refined.

Postscript

That the methodology of constructing CRT's is not as well developed as the technology of their more classical brethren, norm-referenced tests, is apparent. To acknowledge that the field of CR measurement is relatively immature is not, however, to undermine its significance. The methods of CRT development, I feel, will eventually hold a distinguished technical status. Speculative documents on CRT methodology will not, by themselves, make this happen. Indeed, the methodology will improve to the extent that more knowledge of what has been accomplished is produced and made available. The cumulative effects of research, development, and imagination will thus bring us answers to questions such as those posed in this paper. In the meantime, educators can look forward to continued exploration in the field of CR measurement, with the expectation that such efforts will contribute to improved practice in instruction and educational evaluation.

REFERENCES

Baker, E. L. The effects of manipulated item writing constraints on the homogeneity of test items. Journal of Educational Measurement, 1971, 8 (4), 305-309.

Baker, E. L. Beyond objectives: Domain-referenced tests for evaluation and instructional improvement. Educational Technology, 1974, 14, 10-16.

Baker, R. Measurement considerations in instructional product development. In Harris, C. W., Alkin, M. C., & Popham, W. J. (Eds.), Problems in criterion-referenced measurement. CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles, 1974.

Bormuth, J. R. On the theory of achievement test items. Chicago: University of Chicago Press, 1970.

Bortner, M., & Birch, H. G. Cognitive capacity and cognitive competence. American Journal of Mental Deficiency, 1970, 47 (6), 735-744.

Cleary, T. Test bias: Validity of the scholastic aptitude test for Negro and white students in integrated colleges. Research Bulletin 66-31. Princeton, New Jersey: Educational Testing Services, 1966. [ED 018 200]

Cleary, T., & Hilton, T. An investigation of item bias. Educational and Psychological Measurement, 1968, 28 (1), 61-75.

Cole, M., & Bruner, J. Cultural differences and inferences about psychological processes. American Psychologist, 1971, 26, 867-875.

Cronbach, L. J. Test validation. In R.L. Thorndike (Ed.), Educational Measurement, Washington, D. C.: American Council of Education, 1971.

Dahl, T. A. The measurement of congruence between learning objectives and test items. Unpublished doctoral dissertation, University of California, Los Angeles, 1971.

Davis, F. B. & Diamond, J. J. The preparation of criterion-referenced tests. In Harris, C. W., Alkin, M. C. & Popham, W. J. (Eds.), Problems in criterion-referenced testing, CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles, 1974.

Gagne, R. M., & Paradise, N. E. Abilities and learning sets in knowledge acquisition. Psychological Monographs, 75, 1961.

Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.

Glaser, R., & Nitko, A. Measurement in learning and instruction. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.) Washington D. C.: American Council on Education, 1971, 652-670.

Glaser, R., & Resnick, L. B. Instructional psychology. Annual Review of Psychology, 1972, 23, 207-276.

Harris, C. W. Problems of objectives-based measurement. In Harris, C.W., Alkin, M. C., & Popham, W. J. (Eds.), Problems in criterion-referenced measurement. CSE Monograph Series in Evaluation, No. 3, Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles, 1974.

Harris, C. W. Techniques for analyzing test response data. Paper presented at the annual meeting of the American Educational Research Association, Washington, D. C., April, 1975.

Hively, W., & et al. Domain-referenced curriculum evaluation: A technical handbook and a case study from the MINNEMAST Project. CSE Monograph Series in Evaluation, No. 1. Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles, 1973.

Hoepfner, R., Bradley, P. A., Klein, S. P., & Alkin, M. C. CSE/Elementary School Evaluation Kit: Needs Assessment. Boston: Allyn and Bacon, 1972.

Klein, S. P., & Kosecoff, J. Issues and procedures in the development of criterion-referenced tests. ERIC Clearinghouse on Tests Measurement, and Evaluation. New Jersey: Educational Testing Service, 1973.

Millman, J. Criterion-referenced measurement. In Popham, W. J. (Ed.), Evaluation in education: Current applications. Berkeley, California, McCutchan Publishing Corporation, 1974.

Nitko, A. J. Problems in the development of criterion-referenced tests: the IPI Pittsburgh experience. In Harris, C. W., Alkin, M. C., & Popham, W. J. (Eds.). Problems in Criterion-Referenced Measurement, CSE Monograph Series in Evaluation, No. 3, Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles, 1974.

Popham, W. J. Selecting objectives and generating test items for objectives based tests. Paper presented at Conference on Problems in Objectives Based Measurement. Center for the Study of Evaluation, UCLA, 1972.

Popham, W. J. Selecting objectives and generating test items for objectives-based tests. In Harris, C. W., Alkin, M. C., and Popham, W. J. (Eds.). Problems in criterion-referenced measurement, CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, 1974.

Popham, W. J. Educational Evaluation. New Jersey: Prentice Hall, 1975.

Popham, W. J. Tensions and terrors faced by criterion-referenced test developers. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Washington D. C. April, 1975.

Scriven, M. Aspects of curriculum evaluation. In Tyler, R. (Ed.), Perspectives of Curriculum Evaluation. Chicago: Rand McNally, 1967.

Skager, R. Critical characteristics for differentiating among tests of educational achievement. Paper presented at the Annual Meeting of the American Educational Research Association, Washington D.C., 1975.

Smith, E. L. Procedures for generating candidates for learning heirarchies. Paper presented at the annual meeting for Research in Science Teaching. Chicago, April, 1972.

Wilson, H. A. A judgmental approach to criterion-referenced testing. In Harris, C. W., Alkin, M. C., & Popham, W.J. (Eds.) Problems in Criterion-Referenced Measurement. CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles, 1974.